Sensitivity Analysis for Pearson-Lawley Corrections in the Context of Nonignorable Missingness

Guanghan Liu *

Department of Biostatistics

The Johns Hopkins University

Baltimore, MD 21205

Bengt Muthén

Graduate School of Education

University of California, Los Angeles

Los Angeles, CA 90024-1521

August 29, 1994

*This research has been supported by grant N0014-93-1-0619 from the U.S. Navy/Office of Naval Research. The first author should express his deep thanks to the Graduate School of Education of UCLA for hosting him as a statistician and providing very good research facilities.

Abstract

We consider a model sensitivity problem of a dependent variable on several exogenous variables while the dependent variable has some missing data. Under certain assumptions on the model of selected sample and on the selection mechanism, a mixture model is derived and some statistical properties are discussed. This model gives a way to derive Pearson-Lawley (PL) correction formula for the covariance matrix and leads to a modification when the missingness is not ignorable. A sensitivity analysis is then discussed for the PL method. Finally, this modified PL method is applied to a real data set from Project A of Office of Naval Research. The results show some difference from that of using Pearson-Lawley method or of using listwise deletion.

KEY WORDS: sensitivity analysis, nonignorable missingness, Pearson-Lawley (PL) formula, modified PL method.

RUNNING TITLE: Sensitivity Analysis for PL Corrections

1 Introduction

The selection of individuals is common in educational institutions, cooperations, and military organizations. In this situation, a very important issue is to establish a model for a dependent variable like job performance on some exogenous variables like test scores and other background variables for the population such that the prediction and its validity can be studied. Since only those being selected can have measurements of job performance, how to deal with the selection of candidates or missingness of the performance measurements of unselected ones is the problem we want to study in this paper.

In terms of the missingness, there are two basic types of missing mechanism (Little & Rubin, 1987) depending on the relationship of the missingness and the dependent variable of the unselected candidates. One is called missing at random (MAR) or ignorable nonresponse. In this case, a candidate being selected or unselected is independent of his performance measurements but may depend on the exogenous variables. This missingness is ignorable in the sense that estimates can be obtained based on the likelihood function of the observed sample and ignoring the missing observations. The other missing mechanism is called nonignorable missingness, in which case an individual being selected or unselected may depend on the performance measurements.

For the first mechanism, or MAR, many statistical techniques have been proposed (see, e.g., Little & Rubin, 1987; Rubin, 1987). Among those methods, the simplest one is that of using only the observed sample to do statistical inference. This is known as listwise deletion in the missing data literature. In addition to listwise deletion, there are many regression-based adjustment methods like Pearson-Lawley correction, the maximum likelihood procedure, and multiple imputation techniques (Rubin, 1987). All of these methods, except listwise deletion, often give good statistical inference

for MAR cases.

When the data is not missing at random, the situation is more complicated. It is known that methods assuming MAR may often be quite biased when the missingness is nonignorable. Heckman (1976) proposed a selection model which assumed a missing data mechanism in terms of a conditional probability of missing or not missing given the observed measurements. A least-square correction was proposed by Olsen (1980). Lee (1982) gave some approaches using a transformation based on a bias function. Muthén and Jöreskog (1983) pointed out that nonlinearity and heteroscedasticity might occur in nonrandomly selected samples even though the population itself was normally distributed. Recently Little (1994) proposed two unified models for the data and missing mechanism, which include random-coefficient selection model and random-coefficient pattern-mixture model.

It is often necessary to make some assumptions about the missing mechanism for obtaining some statistical inference on a data which involves nonignorable missingness. In this case, a sensitivity analysis is interesting which show how much outcomes are affected by the assumptions of the missing mechanism. Allen, Holland, and Thayer (1994) discussed such a sensitivity analysis for a mixture model (Rubin, 1987) and simplified selection modeling. Brown and Zhu (1994) explored several families of nonignorable missing mechanisms and proposed a compromise solution which provides some protection against nonrandomness.

In this paper, we study the selection problem with one performance variable and several exogenous or covariate variables. A typical example is from military enlistment and job assignment, where the hands-on job performance is of primary concern, while selection is based on test scores and other background variables. In this situation, the Pearson-Lawley correction for the covariance matrix is commonly used for validity assessment.

We first discuss model specification and some statistical properties in Section 2. In Section 3, a modification of the Pearson-Lawley method is derived for the case with nonignorable missingness.

This method gives the same formula as the Pearson-Lawley under the MAR. In Section 4, a sensitivity analysis is presented to show how much outcomes are affected by the assumptions. Some discussion is given on how to choose the unknown parameters according to the possible information. At last, this method is applied to a real data set from Project A of Office of Naval Research (ONR). Comparing with the PL method or the listwise deletion, the modified Pearson-Lawley method gives some different conclusion.

2 Model specification

Suppose y is the performance measurement of interest and $\mathbf{x}$ is a vector of exogenous variables related to selection. Furthermore, we assume that $\mathbf{x}$ is observed always, no missing data occurs; but y is observed only if the individual is selected. Usually there is no information available for y in the unselected sample. Because of this, we may only make a model assumption for the selected sample.

Let R be an indicator of selection such that $R = 1$ if a candidate is selected and y is observed, and $R = 0$ if the candidate is unselected and y is missing.

2.1 A mixture model

Without loss of generality, we assume that the means of y and $\mathbf{x}$ are all 0, otherwise they can be centralized by transformations. Then we assume that when $R = 1$,

$$[y|R = 1] = \mathbf{x}\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \tag{1}$$

for some parameters β and σ^2. This assumes a normal distribution for th conditional distribution of y given $\mathbf{x}$ and when y is observed.

One advantage of only making assumption on the selected sample is that this assumption can be checked since we have observations for both y and $\mathbf{x}$ in the selected sample. Note that the model of (1) may be quite different from assuming a linear model and normality over the whole population as in Heckman (1976), Olsen (1980), Lee (1982), or Muthén and Jöreskog (1983). As Muthén and Jöreskog pointed out, when the whole population follows a linear model with normal residuals, a nonrandom selection procedure results in the model of the selected sample being neither linear nor normal.

Now suppose there is a selection variable s (a latent variable) such that for some function $g(\cdot)$,

$$s = g(\mathbf{x}) + \delta, \quad R = \begin{cases} 1 & \text{if } s \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $g(\mathbf{x})$ contains all contribution to the selection from the exogenous variable $\mathbf{x}$ and δ is a residual term which may be viewed as the contribution from something other than $\mathbf{x}$. This δ is not observed and may depend on both y and $\mathbf{x}$.

Let $[y]$ be the distribution of y. This notation may be a cumulative probability function, or a probability mass function when y is discrete, or a density function when y is continuous. Under the above assumptions, we have a mixture model

$$[y|\mathbf{x}] = [R = 1|\mathbf{x}][y|\mathbf{x}, R = 1] + [R = 0|\mathbf{x}][y|\mathbf{x}, R = 0] \tag{3}$$

where $[y|\mathbf{x}, R = 0]$ is a distribution of y given $\mathbf{x}$ for the unselected sample. This model has been proposed by Glynn, Laird and Rubin (1986) and is named a mixture model. Recently, Allen, Holland and Thayer (1994) applied a similar model to nonignorable nonresponse problems for a

discrete variable y.

With notation as before, let

$$p = P(R = 1|\mathbf{x})$$

be the selection rate for given exogenous variables $\mathbf{x}$, then

$$[y|\mathbf{x}] = p[y|\mathbf{x}, R = 1] + (1 - p)[y|\mathbf{x}, R = 0]$$

With some probability calculation, we have the following result.

Result 1. Let $[y|\mathbf{x}]$, $[y|\mathbf{x}, R = 0]$, $[y|\mathbf{x}, R = 1]$ be corresponding probability mass or density functions. Then

$$[y|\mathbf{x}, R = 0] = \frac{[R = 1|\mathbf{x}]}{[R = 0|\mathbf{x}]} \frac{[R = 0|y, \mathbf{x}]}{[R = 1|y, \mathbf{x}]} [y|\mathbf{x}, R = 1] \tag{4}$$

$$[y|\mathbf{x}] = \frac{[y|\mathbf{x}, R = 1][R = 1|\mathbf{x}]}{[R = 1|y, \mathbf{x}]} = \frac{[y|\mathbf{x}, R = 1]\, p}{[R = 1|y, \mathbf{x}]} \tag{5}$$

When y is a discrete response variable, this result and a proof has been given in Allen, et. al. (1994). A similar argument can be used for the case when y is a continuous variable and leads to the Result 1.

When the missingness is at random, we have $[R = 0|y, \mathbf{x}] = [R = 0|\mathbf{x}]$ and $[R = 1|y, \mathbf{x}] = [R = 1|\mathbf{x}]$. Hence from (4), the distribution of y given $\mathbf{x}$ for the whole population is the same as that of the selected sample. However, for many situations, the missingness may not be ignorable, that is $[R = 0|y, \mathbf{x}]$ may not be independent of y. In this case, we need to know the distribution of y given $\mathbf{x}$ and $R = 0$ given $\mathbf{x}$.

The result of (4) gives a relationship for the distribution of y given $\mathbf{x}$ among those unselected and the distribution of y given $\mathbf{x}$ among those selected. The $[R = 1|y, \mathbf{x}]$ specifies a selection mechanism, or similarly $[R = 0|y, \mathbf{x}]$ is a missing data mechanism, for which we need to make

assumptions. The result of (5) expresses the distribution of y given $\mathbf{x}$ for the whole population as that of the selected sample and the selection mechanism.

After assuming a model for the selection mechanism, the Result 1 will lead us to a model for the full population. A sensitivity analysis will show how much conclusions for the model of $y|\mathbf{x}$ may be affected by varying these assumptions. This is discussed in the later Sections.

2.2 A logistic selection mechanism

Assuming (2), we have

$$[R = 1|y, \mathbf{x}] = P[s > 0|y, \mathbf{x}] = P[\delta > -g(x)|y, \mathbf{x}]$$

Hence, the selection mechanism requires a conditional distribution of residual of the selection variable s after given y and $\mathbf{x}$.

To be explicitly workable, we will take a quadratic logistic model for the conditional distribution of δ given y and $\mathbf{x}$,

$$[\delta > 0|y, \mathbf{x}] = \frac{1}{1 + exp(-\kappa(\lambda_0(\mathbf{x}) + \lambda_1(\mathbf{x})y + \lambda_2(\mathbf{x})y^2))} \tag{6}$$

where $\kappa = \pi/\sqrt{3}$, $\lambda_0(\mathbf{x})$, $\lambda_1(\mathbf{x})$ and $\lambda_2(\mathbf{x})$ are coefficients which may depend on $\mathbf{x}$. Under this assumption, we have

Result 2. If $[\delta|y, \mathbf{x}]$ is given as in (6), then

$$[R = 1|y, \mathbf{x}] = \frac{1}{1 + exp(-\kappa(g(\mathbf{x}) + \lambda_0(\mathbf{x}) + \lambda_1(\mathbf{x})y + \lambda_2(\mathbf{x})y^2))} \tag{7}$$

and

$$E(\delta|y,\mathbf{x}) = \lambda_0(\mathbf{x}) + \lambda_1(\mathbf{x})y + \lambda_2(\mathbf{x})y^2, \tag{8}$$

$$V(\delta|y,\mathbf{x}) = 1. \tag{9}$$

2.3 Statistical Properties of the mixture model

First, let us look at the distribution of $y|\mathbf{x}$ for the whole population. From (5) and (7), we have

$$\begin{aligned}
[y|\mathbf{x}] &= p[y|\mathbf{x}, R=1]/[R=1|y,\mathbf{x}] \\
&= pf(y|\mathbf{x}, R=1) + p\exp(-\kappa(g(\mathbf{x}) + \lambda_0(\mathbf{x}) + \lambda_1(\mathbf{x})y + \lambda_2(\mathbf{x})y^2))f(y|\mathbf{x}, R=1) \\
&= pf(y|\mathbf{x}, R=1) + (1-p)f(y|\mathbf{x}, R=0)
\end{aligned}$$

where

$$f(y|\mathbf{x}, R=0) = \frac{p}{1-p}\exp[-\kappa(g(\mathbf{x}) + \lambda_0(\mathbf{x}) + \lambda_1(\mathbf{x})y + \lambda_2(\mathbf{x})y^2)]f(y|\mathbf{x}, R=1) \tag{10}$$

must be a density function. This requires certain constraints on the parameters of $\lambda_0(\mathbf{x})$, $\lambda_1(\mathbf{x})$ and $\lambda_2(\mathbf{x})$. Without loss of generality, we may vary $\lambda_1(\mathbf{x})$ and $\lambda_2(\mathbf{x})$ but treat $\lambda_0(\mathbf{x})$ as a normalization parameter. Then from (1) and (10), we see that $[y|\mathbf{x}, R=0]$ follows a normal distribution with

$$\text{Mean:} \quad \mu = (\mathbf{x}\beta/\sigma^2 - \kappa\lambda_1(\mathbf{x}))\tau^2 \tag{11}$$

$$\text{Variance:} \quad \tau^2 = (2\kappa\lambda_2(\mathbf{x}) + 1/\sigma^2)^{-1} \tag{12}$$

where $2\kappa\lambda_2(\mathbf{x}) + 1/\sigma^2 > 0$ is a constraint for $\lambda_2(\mathbf{x})$. Therefore, we have the following result.

Result 3. Under the assumptions of (1) and (7), $[y|\mathbf{x}]$ has a mixture distribution resulting from two normal densities.

$$[y|\mathbf{x}] \sim p\, N(\mathbf{x}\beta, \sigma^2) + (1-p)\, N(\mu, \tau^2) \tag{13}$$

where $\lambda_2(\mathbf{x})$ is selected such that $2\kappa\lambda_2(\mathbf{x}) + 1/\sigma^2 > 0$. Moreover, the mean and variance of $[y|\mathbf{x}]$ is given as follows.

$$E(y|\mathbf{x}) = p\,\mathbf{x}\beta + (1-p)\mu \tag{14}$$

$$V(y|\mathbf{x}) = p\,\sigma^2 + (1-p)\,\tau^2 + p(1-p)\,(\mathbf{x}\beta - \mu)^2 \tag{15}$$

It is of interest to look at the correlation between y and the selection variable s given $\mathbf{x}$ (see equations (1) and (2)). This correlation is an important indicator for whether the missingness is ignorable or not. In fact, if the missingness is ignorable, then $s|\mathbf{x}$ is independent of y; otherwise if $cor(y, s|\mathbf{x}) = 0$ then the selection might be unrelated to the dependent variable y and we could expect that the missing is at random. With some calculation, we have the following result.

Result 4. Under the assumption of (1) and (6), the correlation of y and s given $\mathbf{x}$ is

$$\rho = corr(y, s|\mathbf{x}) = cov(y, \delta|\mathbf{x})/\sqrt{V(y|\mathbf{x})V(\delta|\mathbf{x})} \tag{16}$$

where $V(y|\mathbf{x})$ is given at (15) and

$$cov(y, \delta|\mathbf{x}) = \lambda_1(\mathbf{x})V(y|\mathbf{x}) + \lambda_2(\mathbf{x})cov(y, y^2|\mathbf{x})$$

$$V(\delta|\mathbf{x}) = 1 + \lambda_1(\mathbf{x})^2 V(y|\mathbf{x}) + \lambda_2(\mathbf{x})^2 V(y^2|\mathbf{x}) + 2\lambda_1(\mathbf{x})\lambda_2(\mathbf{x})cov(y, y^2|\mathbf{x})$$

and

$$
\begin{aligned}
cov(y, y^2|\mathbf{x}) &= (3p - p^2)\sigma^2\mathbf{x}\beta + (2 - p - p^2)\mu\tau^2 \\
&\quad +p(1-p)[(\mathbf{x}\beta + \mu)(\mathbf{x}\beta - \mu)^2 - \sigma^2\mu - \tau^2\mathbf{x}\beta] \\
V(y^2|\mathbf{x}) &= (3\sigma^4 + 6\sigma^2(\mathbf{x}\beta)^2 + (\mathbf{x}\beta)^4)p + (3\tau^4 + 6\tau^2\mu^2 + \mu^4)(1-p) - \\
&\quad [p(\sigma^2 + (\mathbf{x}\beta)^2) + (1-p)(\tau^2 + \mu^2)]^2
\end{aligned}
$$

Remark 1. The ρ, the correlation between y and s given $\mathbf{x}$, may depend on $\mathbf{x}$ unless $\lambda_1(\mathbf{x})$ and $\lambda_2(\mathbf{x})$ are constants.

3 A modification of the Pearson-Lawley formula

3.1 Pearson-Lawley correction

It is well known that many statistical analyses, such as linear regression, factor analysis, and structural equation modeling can be done using only the mean vector and the covariance matrix without having raw data. In fact, the first two moments give sufficient statistics under the normality assumption. How to get a good estimate of the mean vector and the covariance matrix has received a great deal of attention in statistical literature.

When selection or missing data comes in, how to estimate the mean vector and the covariance matrix is not straightforward. Pearson (1903) and Lawley (1943, 1944) gave adjustment formulas for the mean and covariance matrix for the population after giving the selected sample. Suppose y is the dependent variable which has missing data and $\mathbf{x}$ are the covariates that are observed completely. Let $\mathbf{z} = (\mathbf{x}', y)'$, then under the assumption of MAR or the selection is ignorable, the

PL correction formulas are maximum likelihood estimates without constrains on the mean and covariance matrix of y and $\mathbf{x}$. Let μ^* and Σ^* be the mean vector and the covariance matrix based on the observed sample. Then Pearson-Lawley correction is given as follows.

$$\mu_{\mathbf{z}} = \mu^*_{\mathbf{z}} - \Sigma^*_{\mathbf{zx}}\Sigma^{*-1}_{\mathbf{xx}}(\mu^*_{\mathbf{x}} - \mu_{\mathbf{x}}) \tag{17}$$

$$\Sigma_{\mathbf{zz}} = \Sigma^*_{\mathbf{zz}} - \Sigma^*_{\mathbf{zx}}\Sigma^{*-1}_{\mathbf{xx}}(\Sigma^*_{\mathbf{xx}} - \Sigma_{\mathbf{xx}})\Sigma^{*-1}_{\mathbf{xx}}\Sigma^*_{\mathbf{xz}} \tag{18}$$

If we decompose the matrix according to the size of $\mathbf{x}$ and y and denote

$$\Sigma_{\mathbf{zz}} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

then it is not difficult to find that

$$\begin{aligned} \Sigma_{11} &= \Sigma_{\mathbf{xx}} \\ \Sigma_{21} &= \Sigma^*_{y\mathbf{x}}\Sigma^{*-1}_{\mathbf{xx}}\Sigma_{\mathbf{xx}} \\ \Sigma_{22} &= \Sigma^*_{yy} - \Sigma^*_{y\mathbf{x}}(\Sigma^{*-1}_{\mathbf{xx}} - \Sigma^{*-1}_{\mathbf{xx}}\Sigma_{\mathbf{xx}}\Sigma^{*-1}_{\mathbf{xx}})\Sigma^*_{\mathbf{x}y} \end{aligned}$$

Hence with Pearson-Lawley adjustment, the covariance matrix for the exogenous variables is just the covariance matrix obtained from the total sample. Adjustment is made only on the covariances between $\mathbf{x}$ and y and on the variance of y. In contrast to the analysis based on the observed sample only (listwise deletion), this correction may give significant improvement for the statistical inference.

The adjustment formula (18) can be derived from our model assumption when missingness is ignorable. In fact, when missingness is ignorable, the selection is based on $\mathbf{x}$ and is independent of the performance score y. In the other words, the δ in (2) is independent of y. Then given the

exogenous variable $\mathbf{x}$, the selected sample and the unselected sample follow the same distribution. Under the model of (1), we have

$$[y|\mathbf{x}] = [y|\mathbf{x}, R = 1] \sim N(\mathbf{x}\beta, \sigma^2)$$

Hence after having observations, we can model y for the total population as follows.

$$y = \mathbf{x}\hat{\beta}^* + \epsilon \tag{19}$$

where $\hat{\beta}^*$ is estimated from using the observed sample, $\hat{\beta}^* = \Sigma^{*-1}_{\mathbf{xx}}\Sigma^*_{\mathbf{x}y}$. The ϵ is a random variable which is independent of $\mathbf{x}$ and has mean 0 and variance,

$$\hat{\sigma}^2 = \Sigma^*_{yy} - \Sigma^*_{y\mathbf{x}}\Sigma^{*-1}_{\mathbf{xx}}\Sigma^*_{\mathbf{x}y}$$

Then the covariances between y and $\mathbf{x}$ and the variance of y can be calculated from (19).

Result 5: When the selection is ignorable and $[y|\mathbf{x}, R = 1]$ follows the distribution of (1), we have

1. $\mu_y = \mathbf{x}\hat{\beta}^* + (\mu^*_y - \mathbf{x}^*\hat{\beta}^*)$.
2. $cov(y, \mathbf{x}) = \Sigma^*_{y\mathbf{x}}\Sigma^{*-1}_{\mathbf{xx}}\Sigma_{\mathbf{xx}}$.
3. $cov(y, y) = \Sigma^*_{yy} - \Sigma^*_{y\mathbf{x}}(\Sigma^{*-1}_{\mathbf{xx}} - \Sigma^{*-1}_{\mathbf{xx}}\Sigma_{\mathbf{xx}}\Sigma^{*-1}_{\mathbf{xx}})\Sigma^*_{\mathbf{x}y}$.

which is just the same as the Pearson-Lawley adjustment formula.

We have to note that all the above procedures are based on the assumption of selection being ignorable, that is we have missing at random. This is a crucial condition to derive the Pearson-Lawley formula. Without this assumption, or the missingness not being ignorable, the Pearson-Lawley adjustment may be seriously biased. A modified scheme is proposed as follows.

3.2 A modification of the PL formula

Assume the mixture model of (13), first we shall discuss more about the choice of parameters of $\lambda_1(\mathbf{x})$ and $\lambda_2(\mathbf{x})$. Since both of these parameters are unknown, we have to specify them subjectively. There may be no information about them from observations. Hence, from the point of view of simplicity and plausibility, it is reasonable to first assume that $\lambda_2(\mathbf{x})$ is independent of $\mathbf{x}$. Let

$$v = (2\kappa\lambda_2\sigma^2 + 1)^{-1} \tag{20}$$

Then by (12), $\tau^2 = v\sigma^2$ and

$$\mu = v(\mathbf{x}\beta - \kappa\lambda_1(\mathbf{x})\sigma^2) = v[1 - \kappa\lambda_1(\mathbf{x})\sigma^2/(\mathbf{x}\beta)]\mathbf{x}\beta$$

Now we want to choose a coefficient $\lambda_1(\mathbf{x})$ such that it is proportional to $\mathbf{x}\beta$ and let

$$m = v(1 - \kappa\lambda_1(\mathbf{x})\sigma^2/(\mathbf{x}\beta)) \tag{21}$$

Then this m is also a constant. Under these assumptions, the mean and the variance of the unselected sample is just a scale transformation of the mean and the variance of the selected sample, that is, $\mu = m\,\mathbf{x}\beta$ and $\tau^2 = v\sigma^2$.

In the following discussion, we will use the notations of m and v. Under these, we have $\lambda_1(\mathbf{x})$ $= (v - m)\mathbf{x}\beta/(v\kappa\sigma^2)$ and $\lambda_2 = (1 - v)/(2v\kappa\sigma^2)$. More over, the formulas in Result 3 and Result 4 can be represented by m and v accordingly.

Now we can give a modification of the Pearson-Lawley formula.

Result 6. Under the assumptions of (13), for given $v > 0$ and m, then

$$\begin{aligned} Cov(y, \mathbf{x}) &= (p + (1-p)m)\beta' \Sigma_{\mathbf{xx}} \triangleq f_1(m, p, \sigma^2)\beta' \Sigma_{\mathbf{xx}} \qquad (22) \\ Cov(y, y) &= (p + (1-p)v)\,\sigma^2 \\ &\quad + [(p + (1-p)m)^2 + p(1-p)(1-m)^2]\beta' \Sigma_{\mathbf{xx}} \beta \\ &\triangleq f_2(v, p, \sigma^2)\sigma^2 + f_3(m, p, \sigma^2)\beta' \Sigma_{\mathbf{xx}} \beta \qquad (23) \end{aligned}$$

where $f_1(m, p, \sigma^2)$, $f_2(v, p, \sigma^2)$ and $f_3(m, p, \sigma^2)$ are functions of the selection rate p, the residual variance σ^2 and the parameters m and v. For simplicity, we will denote them by f_1, f_2 and f_3 in the following discussion.

Finally, a modified Pearson-Lawley formula can be obtained by replacing β and σ^2 in (22) and (23) by estimates $\hat{\beta}^* = \Sigma^{*-1}_{\mathbf{xx}} \Sigma^*_{\mathbf{x}y}$ and $\hat{\sigma}^{*2} = \Sigma^*_{yy} - \Sigma^*_{y\mathbf{x}} \Sigma^{*-1}_{\mathbf{xx}} \Sigma^*_{\mathbf{x}y}$. After considering the centering, the mean of y can be estimated as

$$\mu_y = (p + (1-p)m)\mu_x \hat{\beta}^* + \mu^*_y - \mu^*_x \hat{\beta}^* \qquad (24)$$

4 Sensitivity analysis

A sensitivity analysis is of interest to show how much outcomes are affected by assumptions about unknown information and parameters. Such an analysis was performed for a nonignorable nonresponse problem in Allen, Holland and Thayer (1994). Here we will do a sensitivity analysis for the Pearson-Lawley adjustment and its modification to the choices of the parameters m and v.

Suppose that we are interested in a regression analysis of the performance variable y on the exogenous variables $\mathbf{x}$ for the population. The regression coefficients and R^2 can be obtained from

the covariance matrix between y and $\mathbf{x}$. According to the previous discussion, there are three versions of the covariance matrix; one is from using the selected sample only, one is from using the Pearson-Lawley adjustment, and the other is from using the modified PL method.

For example, under the assumption of the selection mechanism (7) and the mixture model (13), an ordinary least-square (OLS) estimator from using the modified PL method would be

$$\hat{\beta}_{mpl} = \Sigma_{\mathbf{xx}}^{-1} Cov(\mathbf{x}, y) = f_1 \hat{\beta}^*$$

where $\hat{\beta}^*$ is an estimate from using the selected sample only and f_1 is a factor defined in (22).

In the following table, we list formulas of variance, covariance, regression coefficients and R^2 for using the selected sample only, using the Pearson-Lawley method and using the modified PL method.

Method	$Cov(\mathbf{x}, y)$	β	$Cov(y, y)$	R^2
Selected sample	$\Sigma^*_{\mathbf{x}y}$	$\hat{\beta}^*$	$\Sigma^*_{yy} \triangleq V_s$	$\hat{\beta}^{*\prime} \Sigma^*_{\mathbf{xx}} \hat{\beta}^* / V_s \triangleq R^2_s$
PL method	$\Sigma_{\mathbf{xx}} \hat{\beta}^*$	$\hat{\beta}^*$	$\hat{\sigma}^{*2} + \hat{\beta}^{*\prime} \Sigma_{\mathbf{xx}} \hat{\beta}^* \triangleq V_{pl}$	$\hat{\beta}^{*\prime} \Sigma_{\mathbf{xx}} \hat{\beta}^* / V_{pl} \triangleq R^2_{pl}$
Modified PL method	$f_1 \Sigma_{\mathbf{xx}} \hat{\beta}^*$	$f_1 \hat{\beta}^*$	$f_2 \hat{\sigma}^{*2} + f_3 \hat{\beta}^{*\prime} \Sigma_{\mathbf{xx}} \hat{\beta}^* \triangleq V_{mpl}$	$f_1^2 \hat{\beta}^{*\prime} \Sigma_{\mathbf{xx}} \hat{\beta}^* / V_{mpl} \triangleq R^2_{mpl}$

4.1 Sensitivity of β and R^2.

Now let us look at the sensitivity properties of the coefficients β and R^2 for various choices of parameters m and v under the model (13).

First, when $m = v = 1$, that is $\lambda_1(\mathbf{x}) = 0$ and $\lambda_2 = 0$, the selection mechanism at (7) depends on $\mathbf{x}$ only. Hence the selection (or missingness) is ignorable. The distribution of the unselected sample is the same as the distribution of the selected sample. In this case, $f_1 = f_2 = f_3 = 1$, and the formulas of (22) and (23) become the same as the PL adjustment formulas. The same statistical inferences can be obtained for β and R^2.

Now suppose $m = 1$. In this case, the mean of $y|\mathbf{x}$ is the same for the selected and the unselected samples. Adjustments may be taken only on the variance of $y|\mathbf{x}$ for the unselected sample. From Result 6, $f_1 = f_3 = 1$, so there is no adjustment on the covariances between y and $\mathbf{x}$. Hence the regression coefficients are the same for both PL adjustment and the modified PL method. The difference of the V_{mpl} and V_{pl} is given by $(f_2 - 1)\hat{\sigma}^{*2}$, which is positive if $v > 1$ and negative if $v < 1$. Hence comparing with R^2_{mpl}, R^2_{pl} is an overestimate if $v > 1$ and an underestimate if $v < 1$.

Another interesting case is when $v = 1$, that is, the variance of the selected sample is the same as the variance of the unselected sample. From Result 6, $f_2 = 1$, the modified PL formula of the covariance between y and $\mathbf{x}$ has a scalar factor of f_1 to the PL covariance. This f_1 is also the scale factor which affects the slope β. It can be seen that this factor is a convex combination of 1 and m with coefficients p and $(1 - p)$, respectively. If m is less than 1, then $f_1 < 1$, the regression coefficients from using the selected sample or using the PL method will be over estimated. On the other hand, when m is larger than 1, the β will be underestimated from those two methods. Since $f_3 = f_1^2 + p(1-p)(1-m)^2$ is always larger than f_1^2, R^2_{mpl} will be smaller than R^2_{pl} if $m < 1$.

Generally, when neither m nor v is one, f_1, f_2 and f_3 are not necessarily one. Similarly as above, f_1 is the scalar factor for the covariance between y and $\mathbf{x}$ and the regression coefficient

β. However, the comparison for R^2_{mpl} and R^2_{pl} is not so clear since R^2_{mpl} depends on both v and m. In Figure 1, we give some contour plots of the relative bias of R^2_{pl} comparing with R^2_{mpl}, i.e. $(R^2_{pl} - R^2_{mpl})$ / R^2_{mpl}, over the parameters m and v after giving p, R^2_{mpl} and σ^2. From these plots, we can see that

1. In general, R^2_{pl} is sensitive to the variation of m and v. For most of the given values of $p = 0.25$ or 0.5 and $R^2_{mpl} = 0.25$ or 0.5, it is quite possible that the relative bias of the R^2_{pl} will be larger than 10% or 20%.

2. When $v > 1$ or $m < 1$, the relative bias is positive, which means the R^2_{pl} is often an overestimate for the population R^2. However, it can be an underestimate when the $v < 1$ or $m > 1$.

3. The relative bias becomes large when the selection rate p is small or the population R^2 is small.

4.2 How to get plausible values for m and v

After seeing the sensitivity of β and R^2 to the values of m and v in the modified PL method, we now discuss on how to obtain some plausible values for the parameters m and v.

First, let us note that for many practical situations, the average score of unselected sample is often smaller than that of selected sample. Hence it is often that $m < 1$. On the other hand, the variance of the performance scores from the unselected sample might be larger than that of the selected sample, that is v will be usually larger than 1. The question is how small this m or how large that v could be.

It is obvious that we need at least two pieces of information to determine the values of m and v. For a given study, people may have some prior information about these two parameters. If

reasonable ranges can be obtained for m and v from experts, then a sensitivity analysis can be given for the regression analysis.

Here is a practical way to obtain the value of v. As in the definition, v is the ratio of the variance of unselected sample to that of selected sample. In many situations, it might be often true that the variation ratio of y is similar to the variation ratio of $\mathbf{x}$. Then the v value can be obtained from the ratios of variances of $\mathbf{x}$ between unselected and selected sample.

For the value of m, it is useful to look at the correlation of y and the selection variable s given $\mathbf{x}$. This is just the correlation of y and the residual δ of regressing s on $\mathbf{x}$. Under the model (13) and the selection mechanism (6), its formula is given at (16). For given $\mathbf{x}$, β and residual variance σ^2, ρ is determined by m and v. Figure 2 gives some contour plots of ρ over various m and v given $p = 0.25$ or 0.5, $\mathbf{x}\beta = 1$ or 2 and $\sigma^2 = 1$. From this Figure, we see that ρ is not very sensitive to v. This is especially so when v is larger than 1. Hence, ρ and m can be roughly determined from one to another. When $m < 1$, ρ is positive; when $m > 1$, ρ is negative. In other words, if the selection residual is positively correlated with the performance score, then the selected sample will have a better average performance score than the unselected sample for given values of $\mathbf{x}$. On the other hand, when the unselected sample has a larger average score, then the residual δ might be negatively related to the performance score y.

For a given data set, p is known, and $\mathbf{x}\beta$ and σ^2 can be estimated from using the selected sample. Then a figure of ρ on m and v can be given as in Figure 2.

Note that ρ is a correlation coefficient of y and δ. The residual δ can be viewed as the contribution from several other exogenous variables which are not used in the selection. Some prior information about those extra exogenous variables might be available from the researchers who design or perform the selection. For example, they may give a rough range for how many are the other variables that are not used in the selection and how much proportion will the other variables contribute to the selection comparing with the variables which are used in the selection.

Remark 2. Since the formula of ρ and V_{mpl} are not simple functions of m and v, it is impossible to solve them for m and v. It is not necessary, however, to get an exact solution for given ρ because both the v and m are only estimates. The contour curves shall be enough to give rough values for the parameters m and v. Moreover, the contour plots will show some sensitivity results.

5 Application to an ONR data set

To illustrate the above approach, we apply this method to a data set of Batch A of the Project A Concurrent Validity Study (see Young, Houston, Harris, Hoffman & Wise (1990)). This data set included Hands-on Job Performance score, 10 subtest scores of the Armed Services Vocational Aptitude Battery (ASVAB) and some other test scores and background variables for nine different jobs. There are 4039 total observations in this dataset. The observations are randomly sampled from the enlisted military persons.

We are interested in a regression analysis for the whole population of the Hands-on job performance score on the ASVAB subtest scores. The 10 ASVAB subtests are Arithmetic Reasoning (AR), Auto & Shop Information (AS), Coding Speed (CS), Electronics Information (EI), General Science (GS), Mechanical Comprehension (MC), Mathematics Knowledge (MK), Numerical Operations (NO), Paragraph Comprehension (PC) and Word Knowledge (WK). For these ASVAB variables, there is a reference population for the selected 4039 sample, which is from the all 650,278 military applicants of 1991 fiscal year. The means and covariance matrix of the ASVAB from the 650,278 applicants is given in Table 1 (a) (see Wolfe et al., (1993)). This sample is taken to be the population from which all the military enlistments are selected and the 4039 Batch A persons are sampled. By doing this, we assume that the distribution of the ASVAB subtest scores for all military applicants will be similar in a consecutive years.

First, Table 1 (b) gives the mean and covariance matrix for the 4039 selected sample. This shall be a consistent estimate for the covariance matrix of ASVAB of the total enlisted population. It can be seen that there is some big difference between the population covariance matrix and this covariance matrix of the selected sample. This implies that the selected (or enlisted persons) are not just a random sample from the applicant population. In fact, the selection is based on the ASVAB subtests and some other variables.

To check the linearity and normality, we perform a regression analysis of y (hands-on job performance score) on the ASVAB for the selected 4039 sample. The residuals of this regression is plotted against the predicted y values on Figure 3. This residual plot looks quite normal. There is no clear violation to the assumption of linearity. Figure 4 gives the histogram and its smooth density curve for those residuals. Although the density curve show a slight skewness to the left, it is still symmetric. So the normality assumption for the residuals might be reasonable.

In order to give a sensitivity analysis, we need assess the values for m and v. Looking at the ratios of variances for the 10 ASVAB variables between the population and the selected sample, we find that they are between 1.19 and 1.47. Since the ratio of the variances between the unselected and selected sample may be smaller than that between the population and the selected sample, we will expect that the values of v is between 1 and 1.47. For the following analysis, we will take the range of v to be from 1 to 1.5.

From the regression analysis of y on the ASVAB variables for the selected sample, we obtain the following information: the residual variance $\sigma^{*2} = 92.05$, the variance of y from the selected sample $V_s = 108.24$, the adjusted mean of y, $\mu_y = \mu_x\hat{\beta}^* + \mu_y^* - \mu_x^*\hat{\beta}^* = 69.77$, and the $\hat{\beta}^{*\prime}\Sigma_{\mathbf{xx}}\hat{\beta}^* = 18.79$. Finally, according to the military enlistment officers, the rate of enlistment to all military jobs is about 33%. This rate can be viewed as the selection rate p since the 4039 observations are randomly sampled from the enlisted population. Then the contour plot of $\rho = corr(y, s|\mathbf{x})$ can be obtained, which is given in Figure 5.

Suppose the prior information about the ρ is that ρ will not be larger than 20% and the correlation is non-negative. Then from the Figure 3, we know that the range for m is between 0.95 and 1.0. With these values of m and v, Table 2 gives several adjusted covariance between Hands-on score and the ASVAB subtest variables. Since the values of m is close to 1, the covariances of y and the ASVAB do not differ a lot. However, the difference of the $var(y)$ between the Pearson-Lawley method and the modified Pearson-Lawley method is apparent.

In the Table 3, the standardized regression coefficients and their t-values are listed. In the last row, the R^2's of the regression are given. We can see that the modified Pearson-Lawley method will have slightly smaller values for the standardized coefficients of β. However, the difference for R^2 is clearer. When $v = 1.5$, the relative difference between the PL method and the modified PL method is about 36%. This is because the modification assumes that the mean of the unselected sample is smaller than the selected sample, while the variance of the unselected sample is larger than the selected sample. Both of these assumptions are quite reasonable for a selection problem like this.

6 Conclusion

In the above study, we considered effects of the nonignorable selection on the Pearson-Lawley adjustment formula for a covariance matrix. The PL formula gives a good correction for the selection bias when the selection is at random, that is, the missingness is ignorable. However, when this condition is not satisfied, the PL formula may be biased. The bias will depend on the model specification and the selection mechanism.

A mixture model with a logistic selection probability is proposed in this paper. Based on this, a modified Pearson-Lawley formula is derived. This gives a further correction to the PL method

when the missingness is not at random. From the sensitivity analysis, we can see to which degree the PL formula will be biased when the selected sample has different mean or variance than that of the unselected sample. For some cases, this bias might be serious. Typically, the relative bias of R^2 for a regression can easily be 20 or 30 percent.

To get the information of the modification parameters, one may assume that the variance ratio of the dependent variable between the population and the selected sample being similar to that of independent variables. This will provide a reasonable range for one of the modification parameters. The other parameter may be accessed from the prior information about the correlation of the performance variable y with the residual of the selection variance s.

Finally, it is very common in practice that the selected and unselected samples have different means and variances. In fact, it is the goal of a selection procedure to choose some special candidates who have better performance ability from a population. Hence it is quite often that a selection will be not at random or the missing is not ignorable. It shall be valuable to have some further investigate for the selection mechanism and use an appropriate modification on the analysis to a data set which involves selection.

References

[1] Allen, N.L. , Holland, P.W. , and Thayer, D. (1994). Approaches to nonignorable nonresponse with application to selection bias. Technical Report, Educational Testing Service.

[2] Brown, C.H. and Zhu, Y. (1994), Compromise solutions to inferences with nonignorable missing data. Draft report, Department of Epidemiology and Biostatistics, University of South Florida.

[3] Heckman, J.J. (1976). The common structure if statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475-492.

[4] Lawley, D. N. (1933-44). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh*: Proceedings, Section A, 62, 28-30.

[5] Lee, L.F. (1982). Some approaches to the correction of selectivity bias. *Review of Economic Studies*, XLIX, 355-372.

[6] Little, R. (1994). Modeling the drop-out mechanism in repeated-measures studies. Draft report, Department of Biostatistics, University of Michigan.

[7] Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

[8] Muthén, B.O. and Gustaffon, J.E. (1994). ASVAB-based job performance prediction

[9] Muthén, B.O. and Gustaffon, J.E. (1994). ASVAB-based job performance prediction and selection: latent variable modeling versus regression analysis. Draft. UCLA.

[10] Muthén, B.O. and Jöreskog, K.G. (1983). Selectivity problems in quasiexperimental studies. *Evaluation Review*, 7, No. 2, 139-174.

[11] Olsen, R.J. (1980). A least squares correlation for selectivity bias. *Econometrica*, 48, No.7, 1815-1820.

[12] Pearson, K. (1903). Mathematical contributions to the theory of evolution-XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London*, Series A, 200, 1-66.

[13] Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

[14] Wolfe, J.H., Alderton, D.L., Larson, G.E. and Held, J.D. (1993). Incremental validity of enhanced computer administered testing (ECAT). Draft. Navy Personnel Research and Development Center.

[15] Young, W.Y., Houston, J.S., Harris, J.H., Hoffman, R.G. and Wise, L.L. (1990). Large-scale predictor validation in Project A: Data collection procedures and data base preparation. *Personnel Psychology*, 43, 301-311.

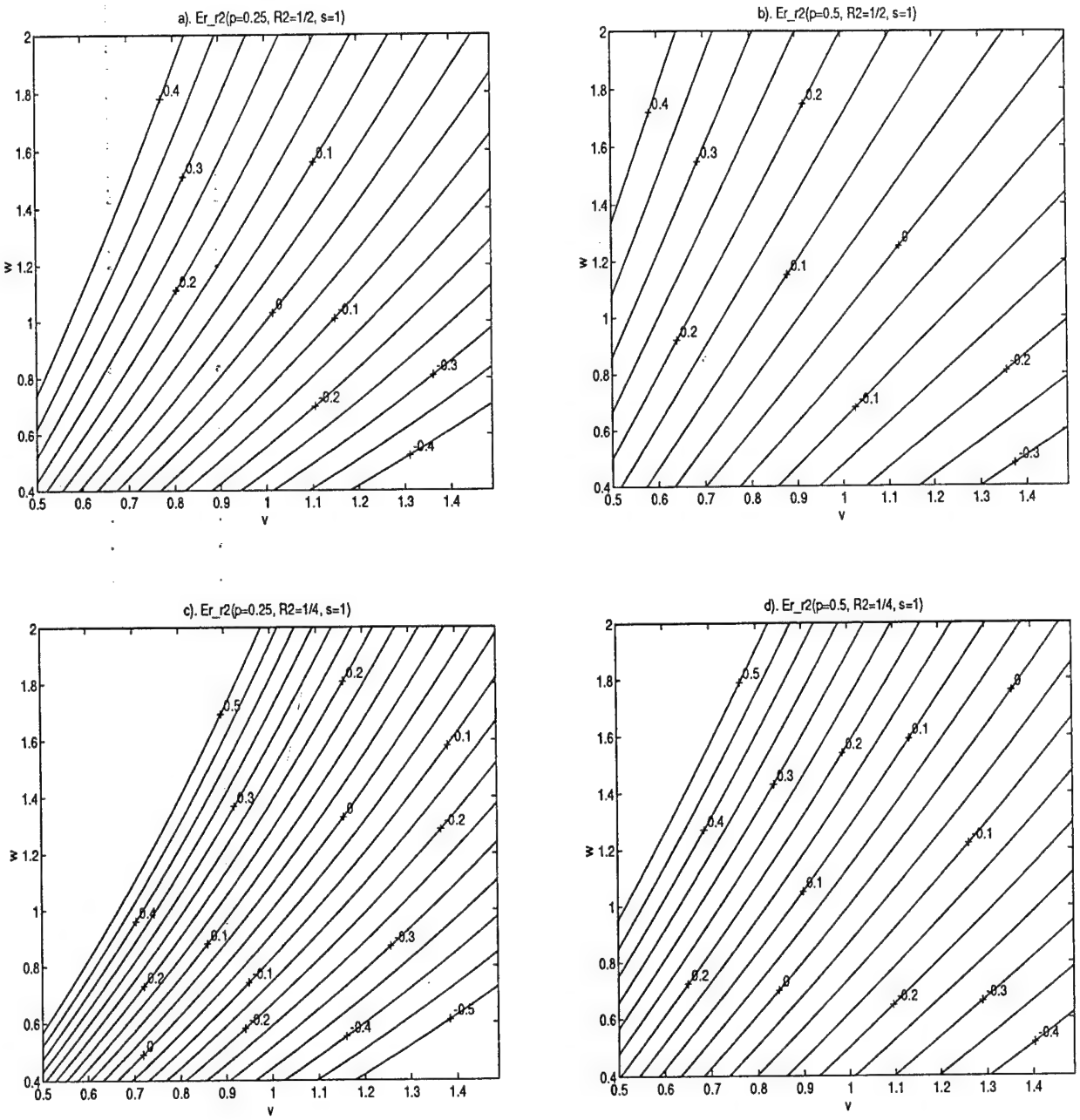

Figure 1: Some contour plots of the relative bias of R^2_{pl} vs R^2_{mpl} on the parameters m and v for given p, R^2_{mpl} and σ^2.

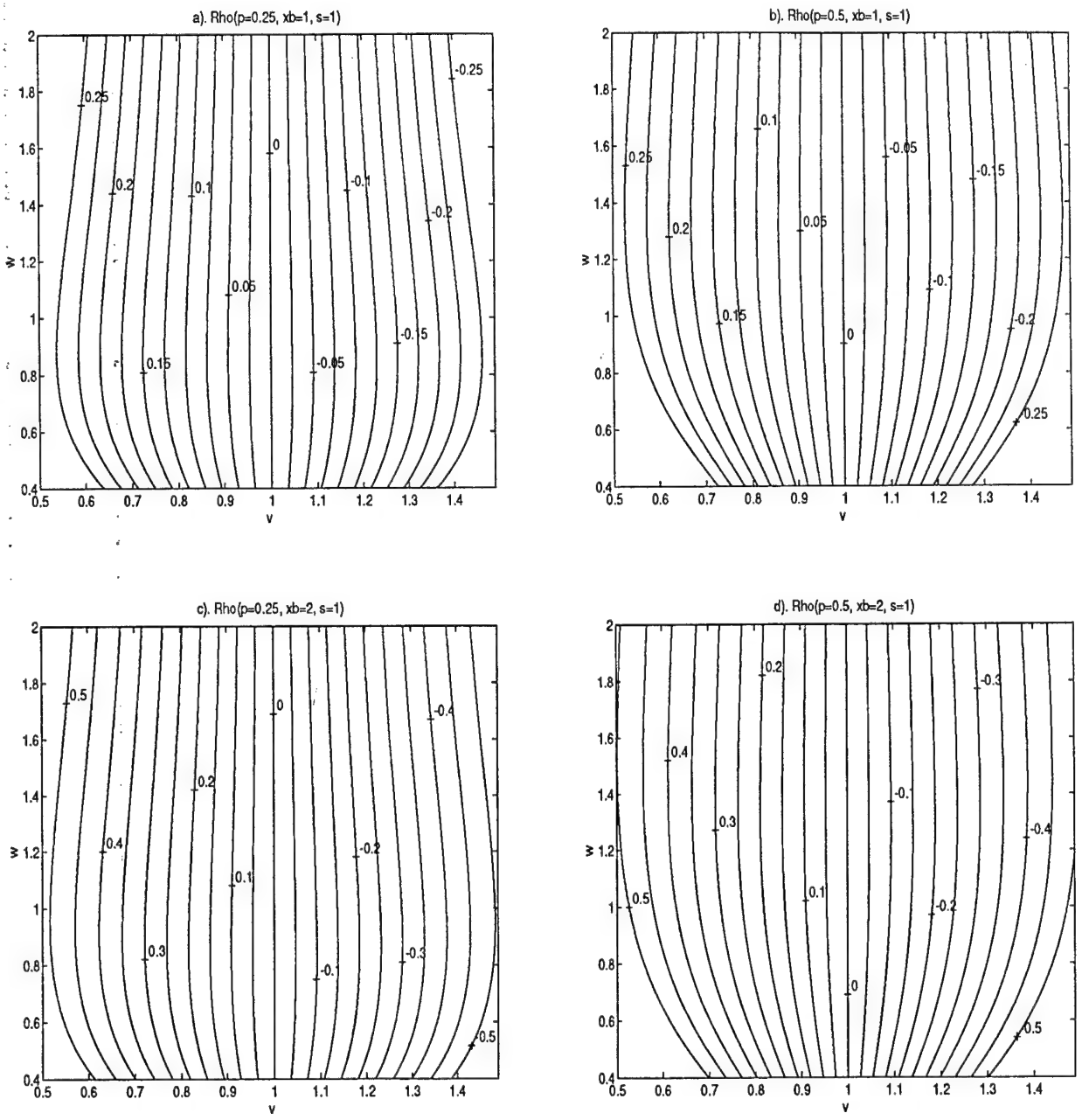

Figure 2: Some contour plots of the correlation $\rho = corr(y,\ s|\mathbf{x}\beta)$ on the parameters m and v for given p, $\mathbf{x}\beta$ and σ^2.

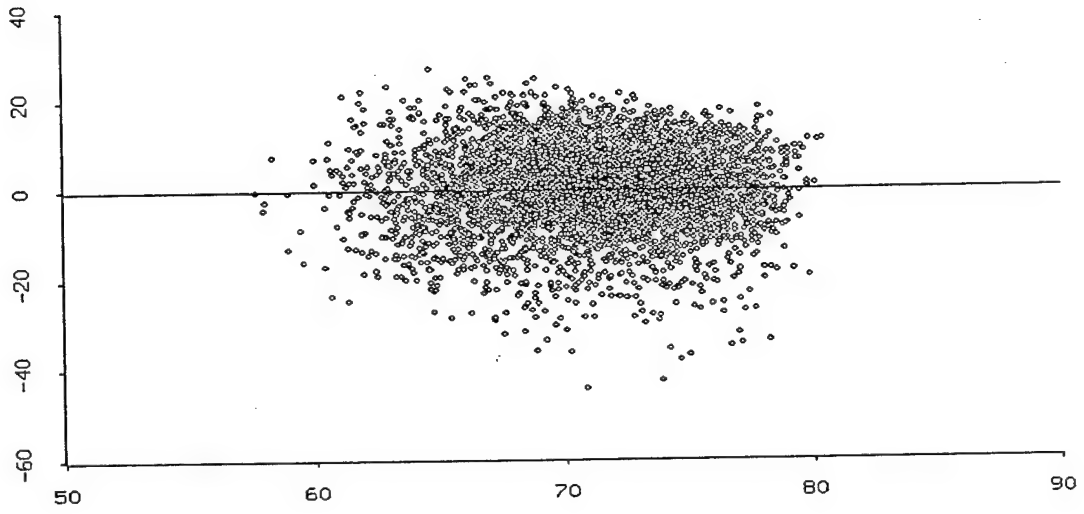

Figure 3: Plot of the residual v.s. predicted y.

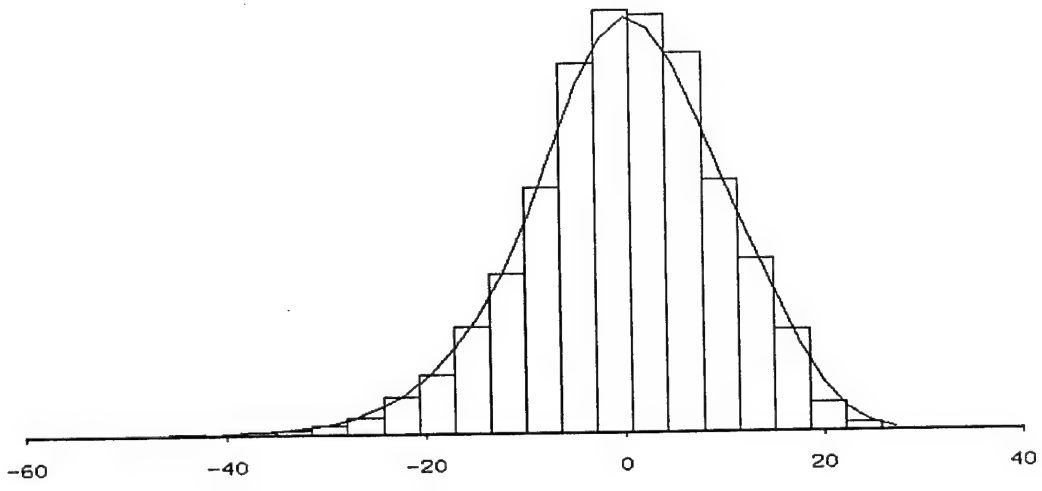

Figure 4: Histogram and the smoothed density curve for the residuals.

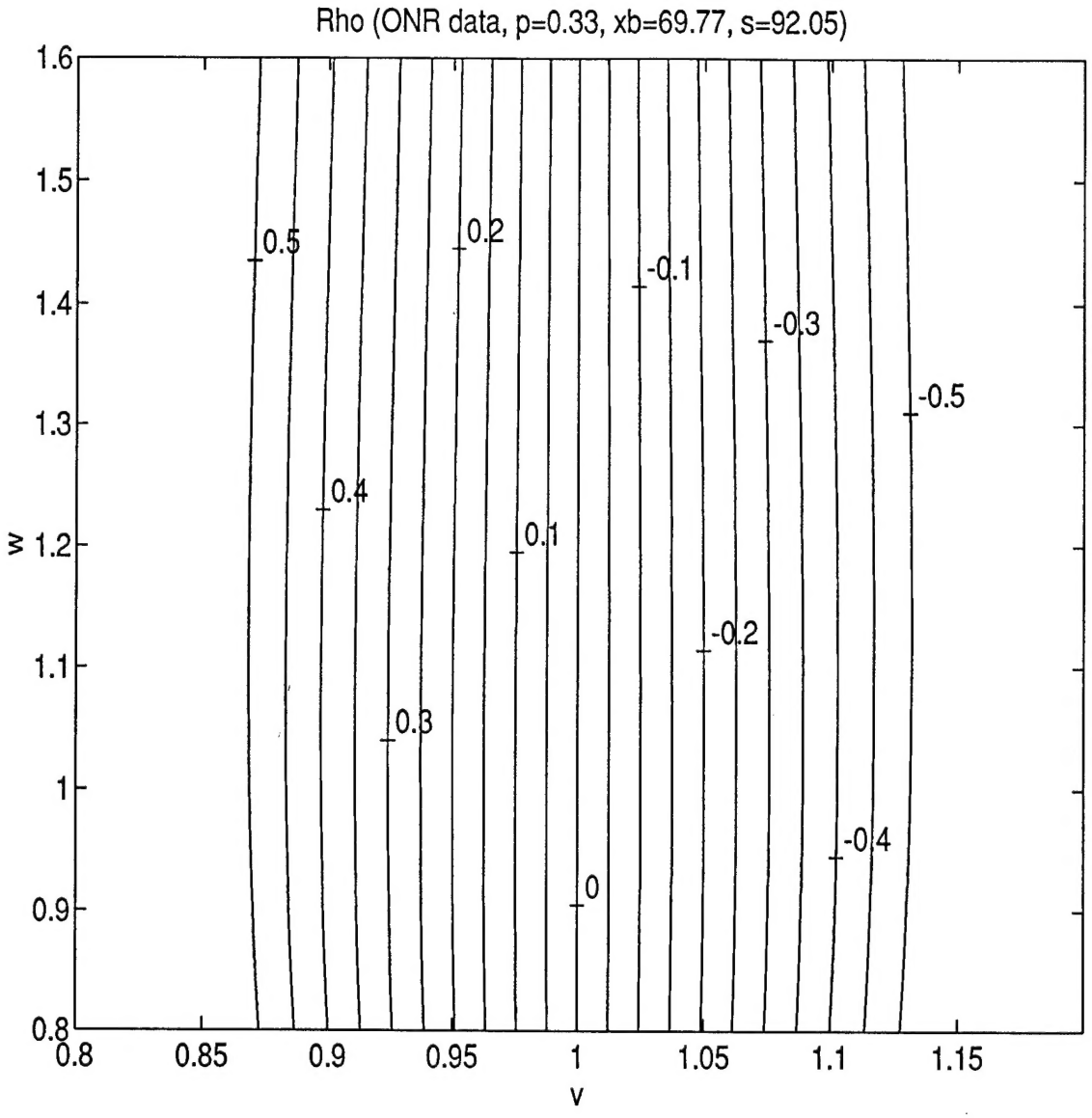

Figure 5: A contour plot of the correlation $\rho = corr(y, s|\mathbf{x}\beta)$ on the parameters m and v for the ONR project A data (n=4039).

Table 1 a). Mean and covariance matrix of the ASVAB from the population sample (n=650,278)*

	AR	AS	CS	EI	GS	MC	MK	NO	PC	WK
AR	74.743	31.736	26.701	37.271	46.346	48.406	53.106	32.583	39.540	37.910
AS	31.736	84.047	4.174	54.333	41.836	51.718	15.659	3.449	24.755	29.437
CS	26.701	4.174	61.025	10.175	17.046	15.769	27.678	40.069	24.006	18.832
EI	37.271	54.333	10.175	78.428	48.520	50.956	28.442	10.307	31.348	34.803
GS	46.346	41.836	17.046	48.520	76.959	51.060	42.246	19.338	42.469	46.458
MC	48.406	51.718	15.769	50.956	51.060	83.306	39.167	16.667	35.271	36.735
MK	53.106	15.659	27.678	28.442	42.246	39.167	75.500	34.543	34.581	31.744
NO	32.583	3.449	40.069	10.307	19.338	16.667	34.543	64.209	25.266	19.116
PC	39.540	24.755	24.006	31.348	42.469	35.271	34.581	25.266	63.426	42.849
WK	37.910	29.437	18.832	34.803	46.458	36.735	31.744	19.116	42.849	54.083
Mean	50.664	51.409	52.266	50.333	50.615	51.941	51.210	52.512	51.156	51.310

* Source: Table A-1 of ECAT Draft report by Wolfe, et. al. (1993).

Table 1 b). Mean and covariance matrix of the ASVAB from the selected sample (n=4039)

	AR	AS	CS	EI	GS	MC	MK	NO	PC	WK
AR	51.548	19.597	8.621	18.701	28.307	27.887	37.579	10.216	20.553	20.925
AS	19.597	70.809	-5.110	36.498	31.474	38.293	11.789	-7.986	15.709	18.073
CS	8.621	-5.110	43.887	-1.333	2.060	-0.282	10.810	22.190	6.799	3.131
EI	18.701	36.498	-1.333	54.846	31.354	34.859	17.237	-4.385	16.573	21.284
GS	28.307	31.474	2.060	31.354	63.783	34.049	28.528	-0.058	29.167	37.200
MC	27.887	38.293	-0.282	34.859	34.049	63.253	26.133	-3.704	18.858	22.077
MK	37.579	11.789	10.810	17.237	28.528	26.133	53.213	13.307	19.818	21.583
NO	10.216	-7.986	22.190	-4.385	-0.058	-3.704	13.307	40.591	2.553	0.043
PC	20.553	15.709	6.799	16.573	29.167	18.858	19.818	2.553	43.046	27.389
WK	20.925	18.073	3.131	21.284	37.200	22.077	21.583	0.043	27.389	43.893
Mean	53.161	54.484	51.661	52.158	51.786	53.467	51.221	52.762	51.787	50.920

Table 2: Covariances and their adjustements between Hands-on performance scores and the ASVAB varaibles for different v and m values

		No_adj	PL	Modified PL		
Para	v		1	1.5	1.5	1
Para	m		1	1	0.95	0.95
factor	f1		1.000	1.000	0.967	0.967
factor	f2		1.000	1.335	1.335	1.000
factor	f3		1.000	1.000	0.935	0.935
HDON	HDON	108.237	110.839	141.674	140.447	109.611
HDON	AR	11.956	17.636	17.636	17.046	17.046
HDON	AS	30.546	35.740	35.740	34.543	34.543
HDON	CS	-1.887	2.567	2.567	2.481	2.481
HDON	EI	19.576	27.316	27.316	26.401	26.401
HDON	GS	16.127	22.023	22.023	21.285	21.285
HDON	MC	24.673	30.906	30.906	29.871	29.871
HDON	MK	11.841	15.773	15.773	15.244	15.244
HDON	NO	-3.592	1.925	1.925	1.860	1.860
HDON	PC	4.868	8.528	8.528	8.242	8.242
HDON	WK	6.073	10.700	10.700	10.341	10.341

Table 3: Standardized regression coefficients, t-values and R^2 for different choices of v and m values

	No_adj		PL		Modified PL					
v			1		1.5		1.5		1	
m			1		1		0.95		0.95	
f1			1		1.000		0.967		0.967	
f2			1		1.335		1.335		1.000	
f3			1		1.000		0.935		0.935	
	Beta	t	Beta	t	Beta	t	Beta	t	Beta	t
AR	-0.027	-1.211	-0.032	-1.312	-0.029	-1.140	-0.028	-1.102	-0.031	-1.269
AS	0.270	13.604	0.291	13.388	0.257	11.627	0.250	11.245	0.283	12.950
CS	0.008	0.478	0.010	0.497	0.009	0.432	0.008	0.418	0.009	0.481
EI	0.034	1.665	0.040	1.773	0.035	1.540	0.034	1.489	0.039	1.715
GS	0.043	1.848	0.047	1.890	0.042	1.641	0.041	1.587	0.046	1.828
MC	0.124	5.872	0.141	6.117	0.125	5.313	0.121	5.138	0.137	5.917
MK	0.109	4.811	0.129	5.669	0.114	4.923	0.110	4.762	0.125	5.484
NO	-0.028	-1.531	-0.034	-1.672	-0.030	-1.452	-0.029	-1.404	-0.033	-1.617
PC	-0.057	-2.929	-0.069	-3.038	-0.061	-2.638	-0.059	-2.552	-0.067	-2.939
WK	-0.098	-4.333	-0.107	-4.252	-0.095	-3.693	-0.092	-3.572	-0.104	-4.113
R^2	0.150		0.146		0.114		0.108		0.138	